Caso de uso: “: Modelo de Redes Neuronales para la Predicción de Supervivencia en Pacientes”

Contexto: En el ámbito de la medicina personalizada y la salud pública, la capacidad de predecir la probabilidad de supervivencia de un paciente en función de sus características clínicas, genéticas, bioquímicas, sociodemográficas y económicas es un objetivo de gran relevancia. Esta predicción permite optimizar la asignación de recursos sanitarios, diseñar estrategias de prevención más efectivas y priorizar intervenciones médicas para los pacientes de mayor riesgo.

Los sistemas de información clínica modernos generan volúmenes masivos de datos heterogéneos que incluyen desde parámetros fisiológicos hasta características del entorno social y económico del paciente. Estos datos, cuando se integran de manera adecuada, pueden alimentar modelos de inteligencia artificial capaces de identificar patrones complejos y no lineales que escapan al análisis tradicional.

En este proyecto, se dispone de un conjunto de datos estructurados que agrupan información detallada de 50.000 pacientes. Cada paciente está descrito mediante múltiples dimensiones, organizadas en tablas independientes: antecedentes clínicos, estilo de vida, biomarcadores en sangre, presencia de mutaciones genéticas relevantes, situación económica y factores sociodemográficos. La variable objetivo es binaria (vive), y representa si el paciente ha superado un cierto periodo crítico (por ejemplo, un año tras diagnóstico o intervención).

Objetivo del Caso de Uso: Diseñar, entrenar y evaluar un modelo de red neuronal que prediga la supervivencia de un paciente (vive = 1) en función de un conjunto multidimensional de variables clínicas, económicas, genéticas y demográficas.

Objetivos específicos:

1. **Integración de datos heterogéneos**. Unir todas las tablas disponibles mediante el identificador único paciente_id, generando un dataset consolidado que preserve la diversidad informacional de cada dimensión (salud, genética, economía, estilo de vida, etc.).
2. **Preprocesamiento avanzado**
    - Codificar adecuadamente las variables categóricas.
    - Normalizar las variables numéricas.
    - Garantizar la calidad de los datos de entrada mediante inspección y tratamiento de valores atípicos.

3. **Diseño del modelo**
    - Construir una red neuronal densa (feedforward) capaz de manejar la complejidad del problema.
    - Ajustar la arquitectura (número de capas, unidades, funciones de activación, dropout, regularización) para maximizar la capacidad predictiva evitando sobreajuste.
4. **Entrenamiento y validación**
    - Entrenar el modelo con un conjunto de entrenamiento y validarlo utilizando un conjunto separado o mediante validación cruzada.
    - Utilizar métricas como **accuracy**, **precision**, **recall**, **F1-score**, y **AUC-ROC** para evaluar su desempeño.
5. **Interpretabilidad y análisis de impacto**
    - Analizar qué variables o grupos de variables (clínicas, económicas, genéticas, etc.) tienen mayor peso en la predicción.
    - Utilizar técnicas como SHAP o LIME para interpretar el modelo y generar explicaciones comprensibles para profesionales médicos.
6. **Comparación con modelos base**
    - Contrastar el desempeño de la red neuronal con otros algoritmos clásicos de clasificación, como regresión logística, árboles de decisión o random forest.
7. **Aplicabilidad real**
    - Establecer escenarios de uso clínico realista donde el modelo pueda integrarse como herramienta de apoyo a la toma de decisiones médicas.

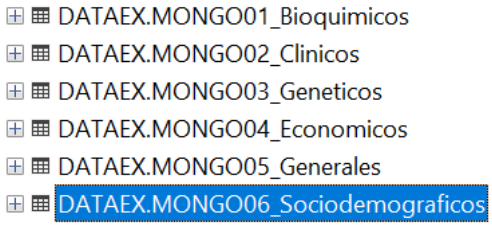
Descripción de las tablas disponibles

1. **Tabla sociodemo**: Contiene variables sociodemográficas del paciente:
    - paciente_id, edad, sexo, estado_civil, nivel_educativo, ocupacion, region, pais_nacimiento, codigo_postal.
2. **Tabla general**: Variables generales relacionadas con el estilo de vida del paciente:
    - paciente_id, fumador, alcohol, actividad_fisica, vive.
3. **Tabla: clínicos**: Condiciones médicas preexistentes o diagnosticadas:

- paciente_id, diabetes, hipertension, obesidad, cancer, enfermedad_cardiaca, asma, epoc.

4. **Tabla bioquímicos**: Indicadores bioquímicos medidos en sangre u otras muestras clínicas:
   - paciente_id, glucosa, colesterol, trigliceridos, hemoglobina, leucocitos, plaquetas, creatinina.
5. **Tabla: genéticos**: Información sobre mutaciones en genes relevantes:
   - paciente_id, mut_BRCA1, mut_TP53, mut_EGFR, mut_KRAS, mut_PIK3CA, mut_ALK, mut_BRAF.
6. **Tabla: económicos**: Variables económicas que pueden influir en el acceso a cuidados médicos:
   - paciente_id, ingresos_mensuales, gastos_salud, seguro_salud, deudas, tipo_empleo, ayudas_publicas.

Todas las tablas se encuentran en azure:



Requisitos para la realización del trabajo:

- Grupos de máximo 1 personas
- Se puede usar cualquier plataforma con python
- Se debe de presentar un documento de máximo 5 diapositivas (sin incluir portadas). Si se presenta más de 5 hojas el trabajo no será evaluado.
- El trabajo debe de contener una solución de viable. No intensificar en la parte técnica.
- El plagio será sancionado con el suspenso automático de la tarea.